Sussex Research Online

The Phi measure of integrated information is not well-defined for general physical systems

Article (Accepted Version)

Barrett, Adam B and Mediano, Pedro A M (2019) The Phi measure of integrated information is not well-defined for general physical systems. Journal of Consciousness Studies, 26 (1-2). pp. 11-20. ISSN 1355-8250

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/81803/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk

The Φ measure of integrated information is not well-defined for general physical systems

Adam B. Barrett*[1], Pedro A. M. Mediano[2]

[1] *Sackler Centre for Consciousness Science* and
Department of Informatics, University of Sussex, Brighton, UK
[2]*Department of Computing*, Imperial College, London, UK

*adam.barrett@sussex.ac.uk (correspondence)

Abstract

According to the Integrated Information Theory of Consciousness, consciousness is a fundamental observer-independent property of physical systems, and the measure Φ of integrated information is identical to the quantity or level of consciousness. For this to be plausible, there should be no alternative formulae for Φ consistent with the axioms of IIT, and there should not be cases of Φ being ill-defined. This article presents three ways in which Φ, in its current formulation, fails to meet these standards, and discusses how this problem might be addressed.

1 Introduction

A key component of integrated information theory (IIT) is the mathematical formalism for supposedly describing quantitatively the extent and nature of the consciousness (subjective experience) generated by any physical system (Oizumi et al, 2014). The theory claims that at every moment that the physical state is updated, there is potential for a conscious experience to be generated. The "intrinsic informational structure" of the mechanisms behind a state transition governs the quality of the experience, whilst the overall quantity of consciousness generated is identical to the system's value of the measure Φ (Tononi et al, 2016). The quantity Φ essentially captures the extent to which the whole system is generating intrinsic information over and above its parts. By *intrinsic* information it is meant that which is independent of the frame of reference imposed by outside observers of the system. The axioms and postulates of IIT state that consciousness is a fundamental, observer-independent property of physical systems, analogous to mass, charge or energy (Tononi and Koch, 2015), and hence imply that Φ is a fundamental physical quantity.

Much of the critique of the Φ measure has been based on the impracticality of its application to empirical neural data, and thus its inability to make testable predictions for IIT (e.g. Bor, 2012). Notably, the computation time required to compute Φ grows faster than exponentially with the number of system components;[1] and it has only ever been computed on a specific kind of toy model

[1]More precisely, the computation time required to compute the effective Φ for a particular system graining grows

system with just a handful of components (Mayner et al, 2017). Here we set testability issues aside, and address the deeper question of whether it it is theoretically possible for Φ to be a fundamental physical quantity. For it to be so, it must be well-defined, and there should be no alternative formulae for Φ consistent with the axioms and postulates of the theory. Here, we list three ways in which it is not well-defined, and hence conclude that further development of the theory and operationalisation of Φ is required.

2 Key quantities for the construction of Φ

This section provides some description of the construction of Φ in words, and writes down the key mathematical quantities (probability distributions) from which it is constructed. A detailed description is not provided; for that the reader is referred to Oizumi et al (2014). Φ has been developed and illustrated via the use of examples of toy model systems consisting of indivisible and discrete binary components (logic gates). These systems evolve in discrete time; at each discrete time-step each component has its state updated according to the specified interactions (mechanisms) present. The dynamics are memoryless (Markovian): the probability distribution for the state at the next time-step only depends on the present state, and not on the past history. Most real complex systems are not easily modelled in this way, and this is a source of the theoretical problems presented below. Defining Φ relies on quantifications of (i) information and (ii) integration, in the system. Each of these components is rehearsed here in turn.

Information is specified as that which the current state of the system contains about a hypothetical past state in which all configurations of the system were *a priori* equally likely.[2] Informally, the concept is that the more past states that are ruled out (or made improbable) by the current state, the greater the information generated. Formally, the key quantity is the joint probability distribution $P_{\rm ce}(\boldsymbol{X}_0, \boldsymbol{X}_1)$ for the states $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ of the system at discrete time $t = 0$ and $t = 1$, given that the system was perturbed at $t = 0$ into all possible states with equal probability (Krohn and Ostwald, 2017). The acronym 'ce' stands for cause-effect. This quantity decomposes as

$$P_{\rm ce}(\boldsymbol{X}_0, \boldsymbol{X}_1) = P(\boldsymbol{X}_1|\boldsymbol{X}_0)P_{\rm u}(\boldsymbol{X}_0)\,, \tag{1}$$

where $P_{\rm u}(\boldsymbol{X}_0)$ is the uniform (or *maximum entropy*) distribution and $P(\boldsymbol{X}_1|\boldsymbol{X}_0)$ is given by the system's dynamics. From this, the conditional distribution $P_{\rm c}(\boldsymbol{X}_0|\boldsymbol{X}_1)$ is extracted:

$$P_{\rm c}(\boldsymbol{X}_0 = \boldsymbol{x}_0|\boldsymbol{X}_1 = \boldsymbol{x}_1) =: \frac{P_{\rm ce}(\boldsymbol{x}_0, \boldsymbol{x}_1)}{\sum_{\boldsymbol{x}^*} P_{\rm ce}(\boldsymbol{x}^*, \boldsymbol{x}_1)}\,. \tag{2}$$

As discussed below: (i) $P(\boldsymbol{X}_1|\boldsymbol{X}_0)$, and hence $P_{\rm ce}(\boldsymbol{X}_0, \boldsymbol{X}_1)$, is only well-defined for Markovian systems; (ii) $P_{\rm u}(\boldsymbol{X}_0)$ is only defined if the set of states is finite (or else compact, i.e. closed and bounded).

Integration is operationalised by comparing probability distributions associated with the whole system to analogous probability distributions associated with a partition of the system. For the

faster than exponentially with the number of components. To obtain the maximum Φ over all grainings is intractable in the absence of a short-cut. See Section 2.1.

[2]Only integrated information of 'cause' is considered here. For integrated information of 'effect' one swaps $t = 0$ and $t = 1$. The final Φ is the minimum of that computed for causes and that computed for effects.

comparison, the probability distributions associated with distinct parts within a partition are taken to be independent. Formally, one computes the distance in probability distribution space between the probability distribution for the whole and the product of the probability distributions for the parts. To define the parts, one must specify a *partition* $\mathcal{P} = \{M^1, M^2, \ldots, M^r\}$ that divides the elements of X into r non-overlapping, non-trivial sub-systems, such that $X = M^1 \cup M^2 \cup \ldots \cup M^r$. With these key elements defined, integration is quantified by considering the distance between

$$P_{\mathrm{c}}(\boldsymbol{X}_0|\boldsymbol{X}_1 = \boldsymbol{x}) \quad \text{and} \quad \prod_k P_{\mathrm{c}}(\boldsymbol{M}_0^k|\boldsymbol{M}_1^k = \boldsymbol{m}^k) \tag{3}$$

where the $\boldsymbol{m}^k$ are the sub-system states corresponding to whole system state $\boldsymbol{x}$ under the partitioning. The greater the distance between these, in probability distribution space, the greater the amount of integrated information (with respect to the given partition). The metric on probability distribution space is taken to be the "earth mover's" (or Wasserstein) distance (Oizumi et al, 2014). Then Φ is the minimum of this distance taken across all possible partitions – in what is commonly known as taking the "cruelest cut" of the system; see Oizumi et al (2014) and Krohn and Ostwald (2017) for details.

2.1 Maximisation over possible grainings

Importantly, to compute Φ, a graining of the system is needed, in space, time and the set of possible states of the components. Since the measure is supposed to be independent of the point of view of the observer, the choice of grainings must be observer-independent. It is prescribed that the grainings to be used are those that lead to the maximum possible value of Φ. Thus, the Φ of a graining is the minimum across partitions of that graining, and the final Φ of the system is the maximum over all possible grainings.

This maximisation over grainings is currently infeasible to carry out in practice for any real physical system, since no compelling short-cuts or approximations yet exist for searching through the infinity of possibilities (Barrett, 2016). Hence computation of Φ is currently intractable for any real physical system. Nevertheless, from the intrinsic (or ontological) perspective, a physical system may instantiate its own maximisation despite that maximisation being infeasible to compute by any external observer. The problems highlighted in Section 3 are distinct from and go beyond this practical computability challenge. We emphasise this issue of graining here since the requirement to maximise over all grainings increases the extent to which Φ is not well-defined, according to Problems 2 and 3 below.

3 Three ways in which Φ is not well-defined

This section lists three ways in which Φ, as currently formulated, is not well-defined.

3.1 Problem 1: There is no canonical metric on the space of states, nor a canonical metric on probability distribution space

In order to compute the earth mover's distance, a metric is required on the space of states, i.e. one requires there to be a well-defined distance between any two states. In IIT-3.0 the Hamming distance

is proposed as the distance when the state of each component of the system is binary. However, for general non-binary states there are a range of possible metrics and no canonical 'intrinsic' choice. For two states $\boldsymbol{x}$ and $\boldsymbol{y}$, example valid expressions for the distance between them are the 'L1 norm' $\sum_i |x_i - y_i|$, the 'L2 norm' $\sqrt{\sum_i (x_i - y_i)^2}$ and the '$L\infty$ norm' $\max_i |x_i - y_i|$. The earth mover's distance is further not the only distance measure for probability distributions. Tegmark (2016) lists several alternatives, and there is no canonical choice. As demonstrated in simulation in Mediano et al (2018), different choices in the construction of a variant Φ measure can lead to profound differences in the behaviour, even on small systems. Therefore, in the absence of a well-argued principle by which to uniquely fix the metrics on the space of states and on probability distribution space, Φ is not well-defined.

3.2 Problem 2: The requirement of a discrete set of states

The maximum entropy distribution on the set of states of the system is not well-defined if the set of states is infinite (except for the case of the set of states being compact, i.e. closed and bounded [Barrett and Seth, 2011]). Thus, for example, the neuron membrane potential can not be taken as the state variable, since there is no way to define precise absolute limits on it. More generally, any system with Gaussian or exponentially distributed state variables does not have a well-defined Φ for this reason. A possible fix might be to only consider grainings into discrete sets of states. However, one would need a canonical method for labelling a discrete set of states obtained from the continuous variable, and to have solved Problem 1 above for there to then be a canonical metric on any discretisation. Then, one would need to show that there always exists an upper bound to the effective Φ over all discrete finite grainings.

3.3 Problem 3: The requirement of Markovian dynamics

Additionally, Φ is not well-defined for a system with non-Markovian dynamics, i.e. one for which the dynamics are not memoryless (Barrett and Seth, 2011). The probability distributions $P(\boldsymbol{X}_1|\boldsymbol{X}_0)$ and $P_{\text{ce}}(\boldsymbol{X}_0, \boldsymbol{X}_1)$ in the formula are not well-defined unless the probability distribution for future states depends only on the current state, and not on the system's past history. This is because for a non-Markovian system the distribution on the past history given $\boldsymbol{X}_0$ is not specified. We highlight that this is an important problem, and not just merely a theoretical construct – brain dynamics are non-Markovian at many levels, ranging from the EEG level (see for example von Wegner et al, 2017), to the level of ionic current fluctuations in membrane channels (see for example Fuliński et al, 1998). More generally, a system may be Markovian with respect to some grainings, but not for all grainings. Given that Φ is supposed to be specified by the maximisation over all possible grainings, (see Section 2.1), it only takes one graining of a given system to have non-Markovian dynamics for Φ to be ill-defined for that system.

For non-Markovian systems, one might perhaps attempt to define Φ as the limit as $k \to \infty$ of the analogous quantity with $\boldsymbol{X}_0$ replaced everywhere with $(\boldsymbol{X}_0, \boldsymbol{X}_{-1}, \boldsymbol{X}_{-2}, \ldots, \boldsymbol{X}_{-k})$, i.e. try setting all past states in an indefinitely long past history to be independent and maximum entropy under the perturbation, and see if there is convergence as the length of past history considered tends to infinity. Such an approach would not however solve the issue for non-ergodic systems: for a non-ergodic system, by definition, there is no convergence of $P(\boldsymbol{X}_1|\boldsymbol{X}_0, \boldsymbol{X}_{-1}, \boldsymbol{X}_{-2}, \ldots, \boldsymbol{X}_{-k})$.

As an example, we consider a non-ergodic system, which has non-Markovian grainings. This system has a variable S (potentially in addition to other variables, which we need not consider to make the point), which follows a random walk:

$$S_{t+1} = S_t + B_t \,, \tag{4}$$

where the B_t are independent identically distributed binary random variables with equal probability of taking the values -1 and 1. Consider the binarisation of this variable S such that the binary state X is given by $X = 1$ if S exceeds some threshold θ, and $X = 0$ otherwise. To compute Φ for this graining we would need the quantity

$$P(X_1 = 1|X_0 = 1) = P(S_1 > \theta|S_0 > \theta) = \sum_{s=\theta+1}^{\infty} P(S_1 > \theta|S_0 = s)P(S_0 = s) \,. \tag{5}$$

But this is not well-defined since there is no well-defined probability density function for S_0: one cannot impose a maximum entropy distribution because the range of values S_0 can take is not a compact set (see Problem 2 above), because as the length of history considered tends to infinity, the set of possible values of S_0 goes to infinity also.

4 Discussion

For IIT to mature as a theory, the three problems above will need to be addressed. This article concludes with some discussion on this.

Problems 2 and 3 arise from needing to quantify the information that the current state holds about some prior state. The maximum entropy distribution is the only possible prior one can impose on the past state, as any other choice will depend on some arbitrary information held by the observer. An empirical distribution can not be used, because not all systems are stationary: the statistics of the system could change the moment any recording is terminated, so one would never know if one has recorded everything that the system could have done. A reformulation of Φ in terms solely of the geometrical and topological structure of the instantaneous state of the system, without reference to past and future states, might be the only way of solving Problems 2 and 3 (Barrett, 2014). This would change the fundaments of the theory somewhat- it would be less about the mechanisms underlying the evolution of the system, and more about simply obtaining a mapping from a physical structure onto the structure of the phenomenal experience associated with the physical structure. Nevertheless, a complex set of mechanisms are needed to generate a system that can exist in multiple complex configurations, so mechanisms would still in some sense be fundamental to consciousness on such an updated theory.

Problem 1 appears to be harder to solve. However one might reformulate the theory, any attempt to create a formula for consciousness as intrinsic information needs to define, spatially, where one system ends and another begins. Without a canonical metric on the space of system configurations, one would not be able to quantify differences between systems and sub-systems in a truly observer-independent fashion. It might be that the possible metrics are heavily constrained by the requirement that the effective Φ must always remain bounded under increasingly fine grainings (see Problem 2); such an investigation is beyond the scope of this paper, but could form the basis for future work.

Successful observer-independent theories for how macroscopic physics emerge from fundamental entities are typically cast in terms of continuous fields, e.g. Einstein's theory of mass and gravitation (general relativity), and Maxwell's theory of electromagnetism (Barrett, 2016). Barrett (2014) proposes that an approach to IIT, and the emergence of consciousness, based on fields might offer advantages over the existing discretization-based approach. It is a debatable supposition that the state of consciousness of a physical system is determined by its structure at a variable spatio-temporal scale and state graining, given by that which happens to maximise Φ for the given system at the given moment (Bayne, 2018). If a formula for the integrated information intrinsic to a field configuration could be obtained, there would be no need to consider alternative grainings of states, or system components. Because human consciousness arises from complex electrical activity in the brain, the hypothesis would be that its fundamental substrate is the integrated information intrinsic to specifically the electromagnetic field (as opposed to say, the gravitational or nuclear force field) it generates (Barrett, 2016); see Barrett (2014) for more on this idea.

Continuing to attempt a formulation of intrinsic information via discrete graining, one might make use of quantities related to Kolmogorov-Sinai (KS) entropy (Sinai, 2009). KS entropy is well-defined for all ergodic systems as a supremum over all grainings. Furthermore, Thurner and Hanel (2012) recently proposed a formalism for defining generalised entropies for non-ergodic systems. Perhaps that could be used to generalise KS entropy to non-ergodic systems, and hence to obtain a universally well-defined intrinsic description of information dynamics.

4.1 Final remarks

We have shown that the supposedly fundamental Φ measure of integrated information, as described in IIT version 3.0 (Oizumi et el., 2014) is not well-defined for general physical systems. We have not addressed here the many variant Φ measures that have been developed for potential practical application to specific classes of systems, see Tegmark (2016) and Mediano et al. (2018) for reviews. These tend to quantify information with respect to the empirical distribution as opposed to the maximum entropy distribution, and can be applied to systems with continuous states (Problem 2 doesn't apply), and moreover to any stationary system (Markovian or not). Further, for a non-linear deterministic system, a distinct approach to operationalising integrated information in terms of topological dimensionality of attractor dynamics has been proposed (Tajima and Kanai, 2017). Any of these measures might be tested for *correlation* with consciousness when computed across choice sets of brain variables (Barrett and Seth, 2011). However, the behaviour of these various measures is very diverse even on small simple networks, so one must remain cautious about considering them as generalisations or approximations of any eventual, 'fundamental' Φ measure (Mediano et al., 2018).

The key idea of IIT, that consciousness is, in some sense, intrinsic information remains intriguing and influential (Tegmark, 2015). However, operationalising this idea and obtaining a candidate universal mathematical description of intrinsic information remains challenging. The current Φ measure is neither universally well-defined, nor fully independent of certain arbitrary choices input into its construction. It is in the best interest of IIT that we recognise and address these problems to move towards a truly plausible measure of phenomenal experience from physical structure.

Acknowledgements

ABB is funded by EPSRC grant EP/L005131/1.

References

Barrett, A.B. (2014) An integration of integrated information theory with fundamental physics, *Front. Psychol.*, 5, 63.

Barrett, A.B. (2016) A comment on Tononi & Koch (2015) 'Consciousness: here, there and everywhere?', *Phil. Trans. R. Soc. B*, 20140198.

Barrett, A.B., & Seth, A.K. (2011). Practical measures of integrated information for time-series data. *PLoS Comput. Biol.*, 7(1): e1001052.

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness, *Neuroscience of Consciousness*, 2018(1), niy007.

Bor, D. (2012) *The Ravenous Brain: How the New Science of Consciousness Explains Our Insatiable Search for Meaning*, New York, NY: Basic Books.

Fuliński, A., Grzywna, Z., Mellor, I, Siwy, Z., & Usherwood, P.N.R. (1998) Non-Markovian character of ionic current fluctuations in membrane channels *Phys. Rev. E* 58, 919.

Krohn, S. & Ostwald, D. (2017) Computing Integrated Information, *Neuroscience of Consciousness*, 2017(1), nix017.

Mayner, W.G.P., Marshall, W., Albantakis, L., Findlay, G., Marchman, R. & Tononi, G. (2017) PyPhi: A toolbox for integrated information theory, *arXiv*, 1712.09644.

Mediano, P.A.M., Seth, A.K., & Barrett, A.B. (2018). Measuring integrated information: Comparison of candidate measures in theory and simulation. *arXiv*, 1806.09373.

Oizumi, M., Albantakis, L. & Tononi, G. (2014) From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0, *PLoS Computational Biology*, 10 (5), e1003588.

Sinai, Y. (2009) Kolmogorov-Sinai entropy, *Scholarpedia*, 4(3):2034.

Tajima, S., & Kanai, R. (2017) Integrated information and dimensionality in continuous attractor dynamics, *Neuroscience of Consciousness*, 2017(1), nix011.

Tegmark, M. (2015) Consciousness as a state of matter, *Chaos, Solitons and Fractals*, 76, 238-270.

Tegmark, M. (2016) Improved Measures of Integrated Information. *PLoS Comput Biol*, 12(11): e1005123.

Thurner, S. & Hanel, R. (2012) The entropy of non-ergodic complex systems - A derivation from first principles. *International Journal of Modern Physics: Conference Series*, 16, 105-115.

Tononi, G. & Koch, C. (2015) Consciousness: here, there and everywhere? *Phil Trans. R. Soc. B* 370, 20140167.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016) Integrated information theory: from consciousness to its physical substrate *Nature Reviews Neuroscience* 17, 450-461.

von Wegner F., Tagliazucchi, E., & Laufs, H. (2017) Information-theoretical analysis of resting state EEG microstate sequences - non-Markovianity, non-stationarity and periodicities. Neuroimage. 158, 99-111.